MATEY: multiscale adaptive foundation models for spatiotemporal physical systems

Pei Zhang[1] **M. Paul Laiu**[2] **Matthew Norman**[3] **Doug Stefanski**[1] **John Gounley**[1]
[1] Computational Sciences and Engineering Division, [2]Computer Science and Mathematics Division,
[3] National Center for Computational Science, Oak Ridge National Laboratory
1 Bethel Valley Road Oak Ridge, Oak Ridge, TN 37831
{zhangp1, laiump, normanmr, stefanskidl, gounleyjp}@ornl.gov

Abstract

Accurate representation of the multiscale features in spatiotemporal systems using vision transformer (ViT) architectures requires extremely long, computationally prohibitive token sequences. To address this issue, we propose an adaptive tokenization scheme which dynamically adjusts the token sizes based on local features. Moreover, we introduce spatiotemporal attention schemes built on axial attention, decoupling full attention into attention in the axial dimensions. We assess the performance of the proposed multiscale adaptive model, MATEY, in a sequence of experiments. The results show that adaptive tokenization is up to eight times more cost-efficient. Compared to a full spatiotemporal attention scheme, we find that decoupled attention requires more training time and larger model sizes to achieve the same accuracy. Finally, we demonstrate in two fine-tuning tasks featuring different physics that models pretrained on PDEBench data outperform the ones trained from scratch, especially in the low data regime with frozen attention.

1 Introduction

Developing foundation models for physical systems is vital for energy generation, earth sciences, and power and propulsion systems. These models offer faster solutions than physics-based simulations and can generalize better across multiple systems than single-purpose AI approaches. However, their application to physical systems, often characterized by multiple sub-processes at different scales, is still in the early stages. The high-resolution solutions of such multiscale multiphysics systems would entail extremely long token sequences, challenging even for advanced supercomputers with existing foundation model algorithms.

Efficient representation of multiscale features in high-resolution inputs has been an active research topic in computer vision. Three broad approaches can be characterized. First, multiscale models like Swin Transformer [Liu+21] and MViTv2 [Li+22] introduce multiple stages with decreasing resolution and increasing feature dimension for efficient hierarchical representations. Second, computational techniques have been developed which facilitate training on long sequences (e.g., sequence parallelism across GPUs [Jac+23]) or reduce the effective sequence length in the attention kernel (e.g., decomposing attention along axial directions [Ho+19]). Third, the actual sequence length can be directly shortened by pruning and merging tokens ([Hau+23; Men+22; Yin+22; BH23]), though this strategy may lead to critical information loss [Liu+24].

These techniques have recently been adopted in sciML for physical systems. For example, the atmosphere foundation model Aurora [Bod+24] uses Swin Transformer, while axial attention is applied by MPP [McC+23]. Despite the progress, computational constraints remain a bottleneck, as existing approaches do not yet handle high-fidelity solutions of applications such as computational

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

fluid dynamics, in which input sequences can easily exceed billions of tokens. More efficient algorithms are needed in foundation models for multiscale multiphysics systems.

In this work, we develop a multiscale adaptive foundation model, MATEY (see Fig. 2), that provides two key algorithmic contributions to address the challenges posed by spatiotemporal physical systems. First, inspired by the adaptive mesh refinement (AMR) technique, we introduce an adaptive tokenization method that dynamically adjusts patch sizes across the system based on local features, which provides up to an $8\times$ reduction in attention cost for similar or higher accuracy. Second, we present a set of spatiotemporal attention schemes based on the axial attention [Ho+19] that differ in their decomposition of long spatiotemporal sequences and identify the cost in time-to-accuracy for axial attention. Finally, we assess the fine-tuning performance of models pretrained on PDEBench [Tak+22] in two out-of-distribution settings, colliding thermals and magnetohydrodynamics (MHD), and observe the pretrained models outperforming random initialized ones.

2 Related work

Scientific foundation models Several research directions have been explored for building foundation models for physical systems, including multiple physics pretraining [McC+23] with PDEBench data, input augmentation with PDE system configurations [Han+24], robust pretraining schemes [Hao+24], fine-tuning effectiveness investigations [Sub+24], and data-efficient multiscale ViT architectures [Her+24]. While these works made remarkable progress, they do not address the issue of token sequence length, which becomes a computation bottleneck when applying ViTs to high dimension or high resolution physical data.

Multiscale ViTs While most multiscale ViTs achieve hierarchical representations via multi-stage attention blocks at different resolutions (e.g., MViTv2 [Li+22] and Swin Transformer [Liu+21]), there are a few focusing on tokenization schemes (e.g., [Yin+22; Fan+24; Zha+24; Hav+23]). One close to our work is the single-stage MSViT with dynamic mixed-scale tokenization [Hav+23], which leverages a learnable gating NN for token refinement controlled via a gate sparsity hyperparameter. It requires careful designing of gate loss functions and adaptive trimming to handle the high overhead cost, which in return hurts gate training. In contrast, our method adaptively adjusts the patch scales directly based on local feature scales, which is simpler and remains effective.

Axial attentions The quadratic scaling nature of attention makes it computationally prohibitive for extremely long token sequences in multidimensional systems. To address this challenge, [Ho+19] proposed the axial attention, which decomposes the full attention into a sequence of attention operations along each axis. It reduces the attention cost from $\mathcal{O}(N^{2d})$ to $\mathcal{O}(N^{d+1})$, for a given d-dimensional system with $N^d = N \times \ldots \times N$ tokens. ViViT [Arn+21] factorized the spatiotemporal attention into spatial- and temporal-dimensions for video classification. [McC+23] applied the axial attention in the Axial ViT (AViT) for spatiotemporal solutions of physical systems. While these spatiotemporal attention schemes can reduce the sequence length and hence the attention cost, their impact on accuracy in physical systems is unclear.

3 Our method

We develop foundation models to predict spatiotemporal solutions of multiple physical systems (Fig. 2). We consider learning a function $\mathbf{u}_{t+t_{\text{lead}}} = \mathbf{f}_{\mathbf{w}}(\mathbf{u}_{t-T+1}, \ldots, \mathbf{u}_t; t_{\text{lead}})$ in a supervised setting to predict the solution at a lead time t_{lead} given a time sequence of T solutions $[\mathbf{u}_{t-T+1}, \ldots, \mathbf{u}_t]$ with $\mathbf{u}_i \in \mathbb{R}^{H\times W\times C}$. The spatial resolution $H \times W$ and the size of physical variables C vary between datasets used in pretraining. For a patch size of $[p_t, p_x, p_y]$, the solution is converted to $N = nt \times npx \times npy$ tokens with $nt = T/p_t, npx = H/p_x,$ and $npy = W/p_y$. Unless otherwise specified, we set $p_t = 1$.

Adaptive tokenization Smaller patch sizes are preferred for better representation accuracy, as ViTs can capture the long-range correlations between patches well but lack inductive biases within patches. However, constant patch sizes that are small enough for good accuracy in physical systems, which often feature multiple scales and exhibit strong spatiotemporal inhomogeneities, result in impractically long token sequence lengths. To address this issue, we propose an adaptive ViT that

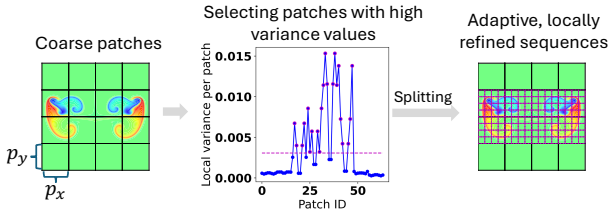

Figure 1: Adaptive tokenization based on local feature variances inside each patch.

dynamically adjusts the tokenization patch sizes according to local physical features. As shown in Fig. 1, to maximize expressiveness, we start with coarse patches, identify the most complex patches in each sample based on a simple metric (i.e., variance of the local features), and further refine the selected patches. Adaptive patch size leads to patches at varying length across samples, which are handled with a padding mask. Patch position and patch area bias are represented following the embedding method in [Bod+24].

AViT, SViT, and ViT In a standard ViT, the attention block learns relationships across the full spatiotemporal patch sequence ($\boldsymbol{Z}^0 = [\boldsymbol{z}_1^0, \ldots, \boldsymbol{z}_N^0]$ with the length $N = nt \cdot nx \cdot ny$ and $\boldsymbol{z}_i^0 \in \mathbb{R}^{C_{\text{emb}}}$). In the simplest setting, the encoder consists of multihead self attention (MHSA) and feed forward multi-layer perceptron (MLP),

$$\widetilde{\boldsymbol{Z}}^\ell = \text{MHSA}(\boldsymbol{Z}^{\ell-1}) + \boldsymbol{Z}^{\ell-1},$$

$$\boldsymbol{Z}^\ell = \text{MLP}(\widetilde{\boldsymbol{Z}}^\ell) + \widetilde{\boldsymbol{Z}}^\ell,$$

with $\ell = 1, \ldots, L$ for L attention blocks. The patch sequence length $N = nt \cdot nx \cdot ny$ in multiscale physical systems is often extremely large, leading to prohibitively high attention costs ($\mathcal{O}(N^2)$). To reduce the spatiotemporal attention cost, various factorized attention mechanisms have been proposed, such as AViT [Ho+19; McC+23] and a spatio-temporal decoupled attention [Arn+21], referred to as SViT here. SViT decouples the full attention into time-attention and space-attention blocks cascaded sequentially, as in

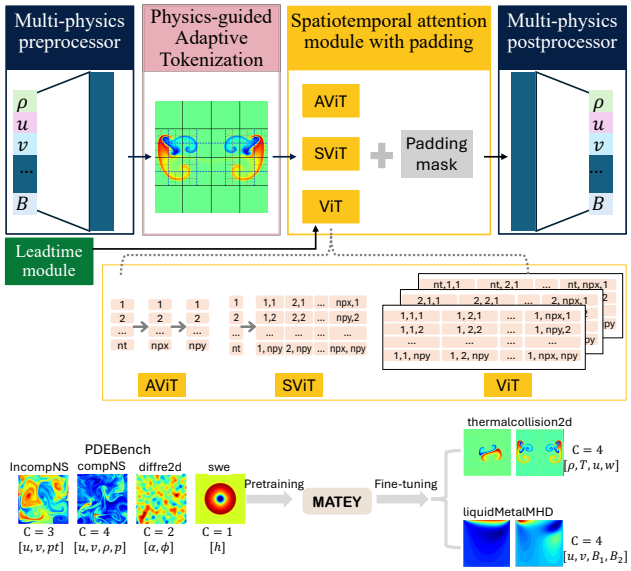

Figure 2: MATEY: multiscale adaptive foundation models for spatiotemporal physical systems.

$$\text{Time sequences:} \quad \boldsymbol{Z}_i^{\ell-1} = \left[\boldsymbol{z}_{(i-1)\cdot nt+1}^{\ell-1}, \boldsymbol{z}_{(i-1)\cdot nt+2}^{\ell-1}, \ldots, \boldsymbol{z}_{(i-1)\cdot nt+nt}^{\ell-1}\right], \quad i = 1, \ldots, nx \cdot ny$$

$$\text{Attention in time:} \quad \widehat{\boldsymbol{Z}}_i^\ell = \text{MHSA}_{\text{time}}\left(\boldsymbol{Z}_i^{\ell-1}\right) + \boldsymbol{Z}_i^{\ell-1}, \quad i = 1, \ldots, nx \cdot ny$$

$$\text{Space sequences:} \quad \widehat{\widehat{\boldsymbol{Z}}}_t^\ell = \left[\widehat{\boldsymbol{z}}_t^\ell, \widehat{\boldsymbol{z}}_{t+nt}^\ell, \ldots, \widehat{\boldsymbol{z}}_{t+nt\cdot(nx\cdot ny-1)}^\ell\right], \quad t = 1, \ldots, nt,$$

$$\text{Attention in space:} \quad \widetilde{\boldsymbol{Z}}_t^\ell = \text{MHSA}_{\text{space}}\left(\widehat{\widehat{\boldsymbol{Z}}}_t^\ell\right) + \widehat{\widehat{\boldsymbol{Z}}}_t^\ell, \quad t = 1, \ldots, nt,$$

$$\text{Feed forward ML:} \quad \boldsymbol{Z}^\ell = \text{MLP}\left(\widetilde{\boldsymbol{Z}}^\ell\right) + \widetilde{\boldsymbol{Z}}^\ell, \quad \ell = 1, \ldots, L,$$

which reduces the attention cost to $npx \cdot npy \cdot \mathcal{O}(nt^2) + nt \cdot \mathcal{O}((npx \cdot npy)^2)$. AViT further decomposes the space-attention in SViT into two axial directions, achieving a cost of $npx \cdot npy \cdot \mathcal{O}(nt^2) + nt \cdot npy \cdot \mathcal{O}(npx^2) + nt \cdot npx \cdot \mathcal{O}(npy^2)$. Due to decoupling, AViT and SViT ignore the spatiotemporal correlations while introducing additional attention blocks which increase the model size. The impact of decoupled attentions on learning efficiency remains unclear. We implement AViT, SViT, and ViT in MATEY and evaluate their performance.

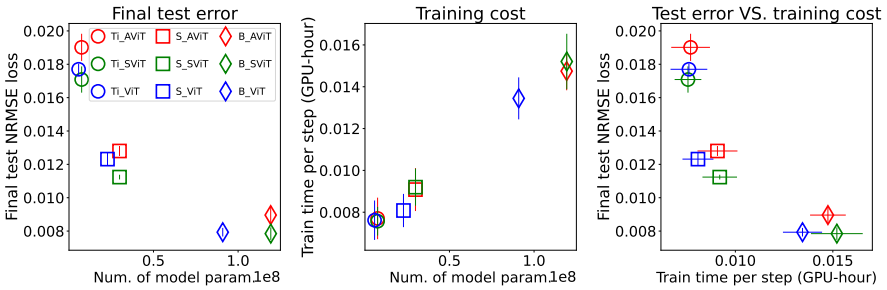

Figure 3: Learning efficiency of three spatiotemporal attention schemes during pretraining in terms of final predictive error and training time cost.

Pretraining and fine-tuning We pretrain the models on five basic 2D systems from PDEBench: incompressible flows, compressible flows, turbulent flows, reaction-diffusion systems, and shallow water equations. We consider two fine-tuning cases: 1) colliding thermals between a cold bubble colliding with a warm bubble from MiniWeather simulations [Nor20] and 2) lid-driven cavity MHD flows [Fam+23]. Training was performed on the Frontier supercomputer at the Oak Ridge Leadership Computing facility, using 92 nodes for pretraining and 4 nodes for fine-tuning.

4 Experiments

We design three experiments to evaluate: 1) the performance of three spatiotemporal attention schemes, AViT, SViT, and ViT, 2) the impact of adaptive tokenization, and 3) the effectiveness of pretrained models on two fine-tuning tasks that feature physics different from the pretraining data.

4.1 Spatiotemporal attention schemes

Fig. 3 compares the training losses, defined as normalized root-mean-square error (NRMSE), against the number of model parameters and training time for the AViT (red), SViT (green), and ViT (blue) schemes. We evaluate these schemes with three model sizes: Tiny (Ti), Small (S), and Base (B) with 3, 6, and 12 heads and hidden dimensions of 192, 384, and 768, respectively [Tou+22]. In general, we find that SViTs and ViTs are more computationally- and representation-efficient than AViTs, in that they achieve lower losses with the same training time and smaller model sizes.

4.2 Adaptive tokenization

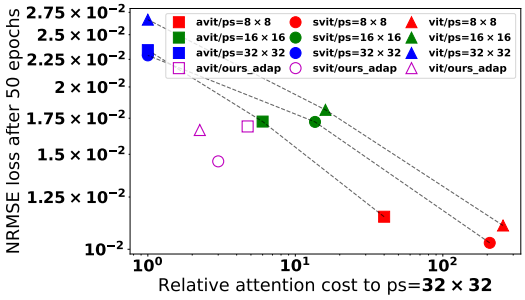

Figure 4: NRMSE loss for attention schemes with adaptive tokenization (ours_adap) and constant patch sizes. Cost is estimated from sequence length. Patch size/resolution represents grid of pixels per patch.

In Fig. 4, we evaluate our adaptive tokenization method coupled with AViT, SViT, and ViT, together with three patch resolutions: ps=8×8, 16×16, and 32×32. For the same attention scheme, as expected, increasing resolution leads to lower error but also substantially increases the training cost. Our adaptive scheme, which starts with 32×32 and locally refines to 8×8 on selected patches, achieves comparable or better accuracy than uniform 16×16 patches despite reduced training costs. The cost reduction depends significantly on the spatiotemporal attention, being almost $8\times$ more efficient than constant patch sizes for ViT.

4.3 Effectiveness of pretraining in colliding thermals and MHD fine-tuning tasks

By comparing the test errors after fine-tuning between pretrained and randomly initialized models, we aim to assess four questions:

1. Does pretraining improve accuracy with limited data?

2. Is pretraining still useful when the downstream tasks have a distinct set of physical variables?
3. How does limited fine-tuning (freezing attention and training the preprocessor/postprocessor only ['PREPOST']) compare to full fine-tuning ['ALL']?
4. How does fine-tuning data size affect convergence?

For the thermal collision dataset, Fig. 5 compares the test loss with PREPOST using pretrained (i.e., '*_pretrain') and randomly initialized models (i.e., '*_INIT') for different training data sizes ranging from one set of thermal collision time-trajectories to 96 sets of trajectories. Pretrained models achieve significantly lower error than starting from scratch with randomly initialized weights. With increasing fine-tuning data, test errors of pretrained and randomly initialized models converge to different values. The lower converged error from pretrained models suggests attention blocks clearly learn transferable knowledge from pretraining.

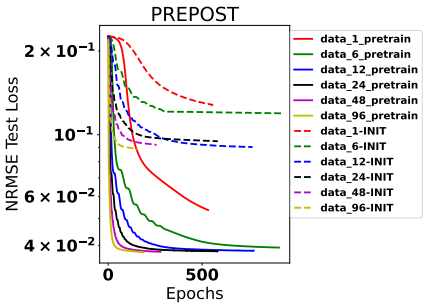

Figure 5: NRMSE loss for test set at different training data sizes in fine-tuning of colliding thermals.

Fig. 6 shows the final test NRMSE errors after fine-tuning against data sizes, from pretrained models ('PREPOST', 'ALL') and from scratch ('*_INIT') for liquid metal MHD in lid-driven cavity flows. Pretrained models achieve lower fine-tuning errors, similar to colliding thermals. Regarding the two fine-tuning strategies, the advantage of pretraining vanishes with increasing data for 'ALL' but persists for 'PREPOST'. This is a result of model expressibility, training data size, and the similarity between training and testing tasks. Models with limited expressibility, such as 'PREPOST*' with its attention blocks frozen, consistently show an accuracy gap, even with more training data, as they cannot fully represent the data complexity. In contrast, highly expressive models (i.e., 'ALL*' with all parameters trainable) can capture all training data information when trained on limited data but often show high test errors; as more training data is provided, they generalize better and lead to an improved test error. In our fine-tuning, the randomly initialized models perform well in testing even with a single data configuration (equivalently, 1989 samples), likely due to the similarity between training and testing tasks. Future work will explore more challenging scenarios.

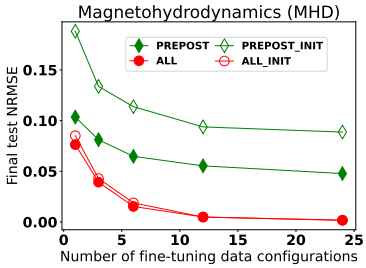

Figure 6: NRMSE loss for test set against training data sizes in fine-tuning of liquid metal MHD.

5 Conclusions

In this paper, we make three contributions that will advance the development of foundation models for multiscale physical systems. First, we find that while some data efficiency is lost in a fully decoupled spatiotemporal attention scheme such as AViT, SViT provides an intriguing balance of computational and data efficiency versus the standard ViT approach. Yet using SViT alone does not sufficiently address the computational challenges associated with attention for high spatial resolutions. Second, we instead suggest that our adaptive tokenization scheme provides a promising approach for working with high resolution data. Adaptivity has the potential to be sufficiently flexible and expressive to represent the dynamic and sparse nature of the multiscale features in physical data. Third, we suggest an alternative path to evaluate foundation models for multiscale physical systems that focuses on fine-tuning problems involving out-of-distribution physics governed by different equations with distinct sets of physical variables. In two such settings, colliding thermals and magnetohydrodynamics, we find that while pretraining does provide an advantage, its impact is much more muted compared to fine-tuning on the same set of variables, suggesting that additional effort is required to obtain truly foundational models in this space.

Acknowledgments

This research is sponsored by the Artificial Intelligence Initiative as part of the Laboratory Directed Research and Development (LDRD) Program of Oak Ridge National Laboratory, managed by

UT- Battelle, LLC, for the US Department of Energy under contract DE-AC05-00OR22725. This material is based in part upon work carried out in the 'Center for Simulation of Plasma - Liquid Metal Interactions in Plasma Facing Components and Breeding Blankets of a Fusion Power Reactor' project, supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research and Office of Fusion Energy Sciences, Scientific Discovery through Advanced Computing (SciDAC) program. This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DEAC05-00OR22725.

References

[Arn+21] Anurag Arnab et al. *ViViT: A Video Vision Transformer*. 2021. arXiv: 2103.15691 [cs.CV]. URL: https://arxiv.org/abs/2103.15691.

[BH23] Daniel Bolya and Judy Hoffman. "Token Merging for Fast Stable Diffusion". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2023, pp. 4599–4603.

[Bod+24] Cristian Bodnar et al. "Aurora: A foundation model of the atmosphere". In: *arXiv preprint arXiv:2405.13063* (2024).

[Fam+23] F. Fambri et al. "A well-balanced and exactly divergence-free staggered semi-implicit hybrid finite volume / finite element scheme for the incompressible MHD equations". In: *Journal of Computational Physics* 493 (2023), p. 112493. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2023.112493. URL: https://www.sciencedirect.com/science/article/pii/S0021999123005880.

[Fan+24] Qihang Fan et al. *ViTAR: Vision Transformer with Any Resolution*. 2024. arXiv: 2403.18361 [cs.CV]. URL: https://arxiv.org/abs/2403.18361.

[Han+24] Zhou Hang et al. "Unisolver: PDE-Conditional Transformers Are Universal PDE Solvers". In: *arXiv preprint arXiv:2405.17527* (2024).

[Hao+24] Zhongkai Hao et al. "DPOT: Auto-regressive denoising operator transformer for large-scale PDE pre-training". In: *arXiv preprint arXiv:2403.03542* (2024).

[Hau+23] Joakim Bruslund Haurum et al. "Which Tokens to Use? Investigating Token Reduction in Vision Transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. Oct. 2023, pp. 773–783.

[Hav+23] Jakob Drachmann Havtorn et al. "MSViT: Dynamic Mixed-scale Tokenization for Vision Transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 838–848.

[Her+24] Maximilian Herde et al. "Poseidon: Efficient Foundation Models for PDEs". In: *arXiv preprint arXiv:2405.19101* (2024).

[Ho+19] Jonathan Ho et al. "Axial attention in multidimensional transformers". In: *arXiv preprint arXiv:1912.12180* (2019).

[Jac+23] Sam Ade Jacobs et al. *DeepSpeed Ulysses: System Optimizations for Enabling Training of Extreme Long Sequence Transformer Models*. 2023. arXiv: 2309.14509 [cs.LG]. URL: https://arxiv.org/abs/2309.14509.

[Li+22] Yanghao Li et al. "Mvitv2: Improved multiscale vision transformers for classification and detection". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 4804–4814.

[Liu+21] Ze Liu et al. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[Liu+24] Yixin Liu et al. *Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models*. 2024. arXiv: 2402.17177 [cs.CV]. URL: https://arxiv.org/abs/2402.17177.

[McC+23] Michael McCabe et al. "Multiple physics pretraining for physical surrogate models". In: *arXiv preprint arXiv:2310.02994* (2023).

[Men+22] Lingchen Meng et al. "Adavit: Adaptive vision transformers for efficient image recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 12309–12318.

[Nor20] Matthew R Norman. *miniWeather*. Tech. rep. Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), 2020.

[Sub+24] Shashank Subramanian et al. "Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior". In: *Advances in Neural Information Processing Systems* 36 (2024).

[Tak+22] Makoto Takamoto et al. "PDEBench: An extensive benchmark for scientific machine learning". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 1596–1611.

[Tou+22] Hugo Touvron et al. "Three things everyone should know about vision transformers". In: *European Conference on Computer Vision*. Springer. 2022, pp. 497–515.

[Yin+22] Hongxu Yin et al. "A-vit: Adaptive tokens for efficient vision transformer". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10809–10818.

[Zha+24] Enzhi Zhang et al. *Adaptive Patching for High-resolution Image Segmentation with Transformers*. 2024. arXiv: 2404.09707 [cs.CV]. URL: https://arxiv.org/abs/2404.09707.